IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Martin Ng>
<16/6/2024>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- In this capstone project, we aim to predict the successful landing of the SpaceX Falcon 9 first stage using various machine learning classification algorithms.
- The key steps involved in this project are:
  - Collecting, processing, and formatting the data
  - Conducting exploratory data analysis
  - Creating interactive data visualizations
  - Applying machine learning techniques for prediction
- Our analyses reveal that certain features of the rocket launches are correlated with the outcome, whether it's a successful or failed landing.
- Based on the findings, the decision tree algorithm appears to be the most suitable machine learning model for predicting the successful landing of the Falcon 9 first stage.

Introduction

- In this capstone project, we aim to predict whether the Falcon 9 first stage will land successfully. This is an important consideration, as SpaceX advertises Falcon 9 rocket launches on its website at a cost of $62 million, while other providers charge upwards of $165 million per launch. A significant portion of the savings is due to SpaceX's ability to reuse the first stage. Therefore, by accurately predicting the success of the first stage landing, we can estimate the true cost of a Falcon 9 launch.
- This information could be valuable for other companies that want to bid against SpaceX for rocket launch contracts. It's worth noting that most unsuccessful landings by SpaceX are actually planned, as the company sometimes performs controlled landings in the ocean.
- The main question we are trying to answer in this project is: Given a set of features about a Falcon 9 rocket launch, such as payload mass, orbit type, launch site, and so on, can we accurately predict whether the first stage of the rocket will land successfully?

Section 1

Methodology

Methodology

Executive Summary

- Data collection and wrangling using SpaceX API and Web scraping
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-nearest Neighbors

Data Collection and Wrangling

- The dataset was subsequently processed to address any missing data points. Additionally, the categorical features were encoded using the one-hot encoding technique.
- Furthermore, a new column named 'Class' was added to the dataset. This column indicates whether a given launch was successful (denoted by '1') or failed (denoted by '0').
- Finally, the final dataset comprises 90 rows (or instances) and 83 columns (or features).

Data Collection – SpaceX API

- SpaceX API provides numerous types of rocket launches by SpaceX.
- Data is filtered to contain only Falcon 9 launches
- All missing values in the data were replaced with the mean value of the corresponding column
- The final dataset contains 90 rows and 17 columns. The image below shows the first 5 rows of the dataset

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

Data Collection - Scraping

- The website contains the data of Falcon 9 launches
- The final dataset contains 121 rows and 11 columns. The image below shows the first 5 rows of the dataset:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

EDA with Data Visualization

- Pandas and Numpy
  - To derive basic insights about the collected dataset, functions from the Pandas and NumPy libraries were utilized. This analysis included:
    - Determining the number of launches from each launch site
    - Identifying the frequency of occurrence for each type of orbit
    - Calculating the number and occurrence of each mission outcome (i.e., successful or failed launches).
- Matplotlib and Seaborn
  - To further explore the data, functions from the Matplotlib and Seaborn data visualization libraries were utilized. This allowed the creation of scatterplots, bar charts, and line charts to visually represent the data. These visualizations were used to investigate the relationships between several features, including:
    - The relationship between flight number and launch site
    - The relationship between payload mass and launch site
    - The relationship between success rate and orbit type

EDA with SQL

- The space mission data is analyzed using SQL queries to extract and provide answers to various questions, including:
  - Identifying the distinct launch sites where the space missions have taken place.
  - Calculating the total payload mass that has been carried by boosters launched under the NASA (CRS) program.
  - Determining the average payload mass carried by the booster version 'F9 v1.1'.
- In essence, the SQL queries are used to explore and gain insights from the space mission data by focusing on aspects such as launch locations, payload capacities for specific agencies, and performance characteristics of different booster versions.
- The goal is to leverage the data and extract valuable information to better understand and analyze the space missions through these targeted SQL-based queries.

Build an Interactive Map with Folium

- The Folium library is utilized to create interactive visualizations of the space mission data through the use of maps. Specifically, the Folium library is employed to:
  - Plot all the launch sites on a map, providing a spatial representation of where the space missions have originated.
  - Distinguish between the successful and failed launches for each launch site on the interactive map, allowing for a clear visual differentiation of the mission outcomes.
  - Indicate the distances between each launch site and its closest proximity, such as the nearest city, railway, or highway. This added contextual information helps to better understand the geographic landscape and accessibility of the launch locations.
- In summary, the Folium library enables the creation of dynamic, map-based visualizations that help to spatially represent and analyze various aspects of the space mission data, including launch site locations, mission success rates, and the surrounding infrastructure and geography.

Build a Dashboard with Plotly Dash

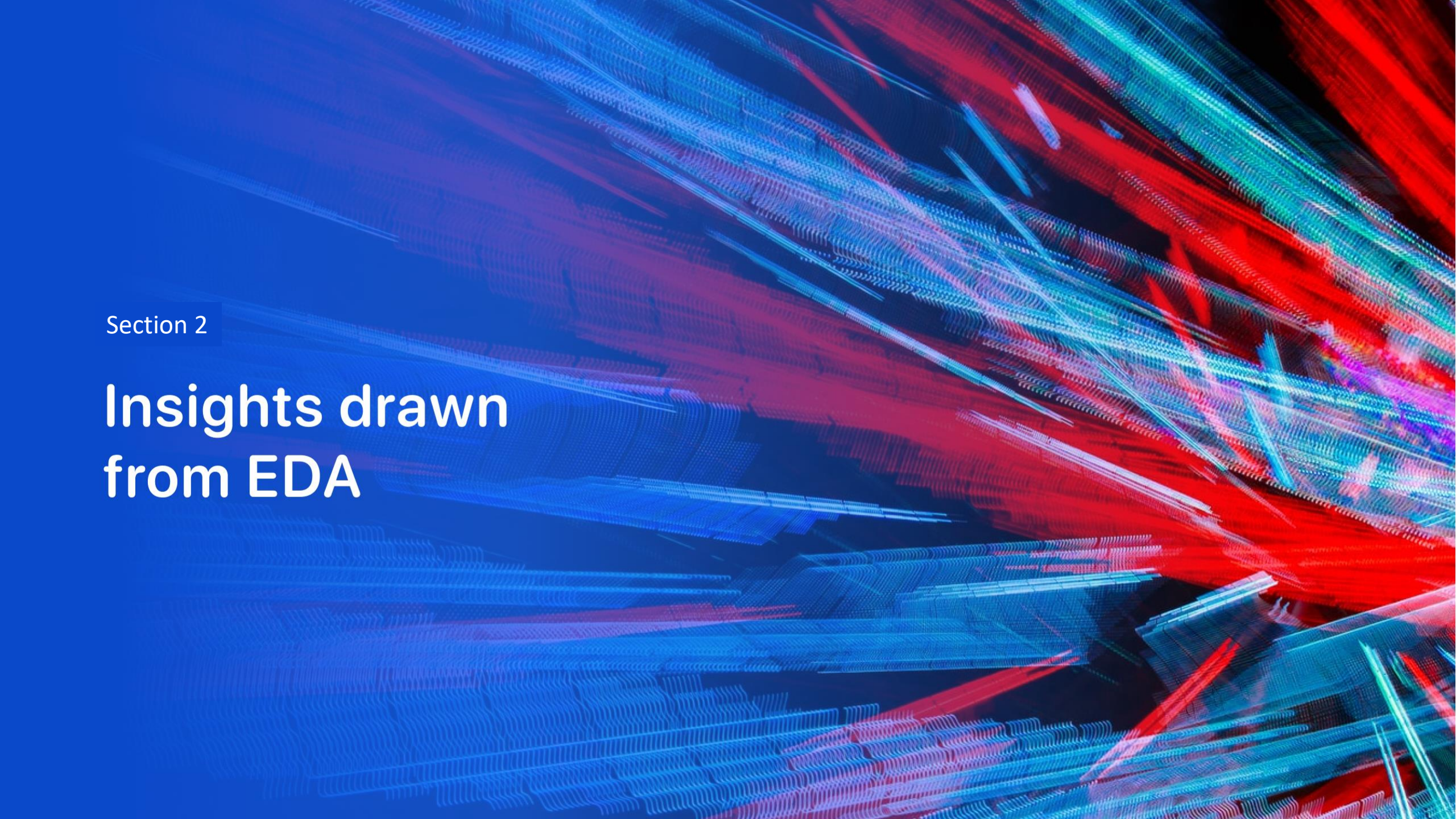
- The interactive site built using Dash allows users to toggle inputs via a dropdown menu and range slider. This site presents:
  - A pie chart showing the total successful launches from each launch site.
  - A scatterplot illustrating the correlation between payload mass and mission outcome (success or failure) for each launch site.
- The Dash-powered interface enables users to dynamically explore the space mission data, with the visualizations reflecting changes to the input parameters.

Predictive Analysis (Classification)

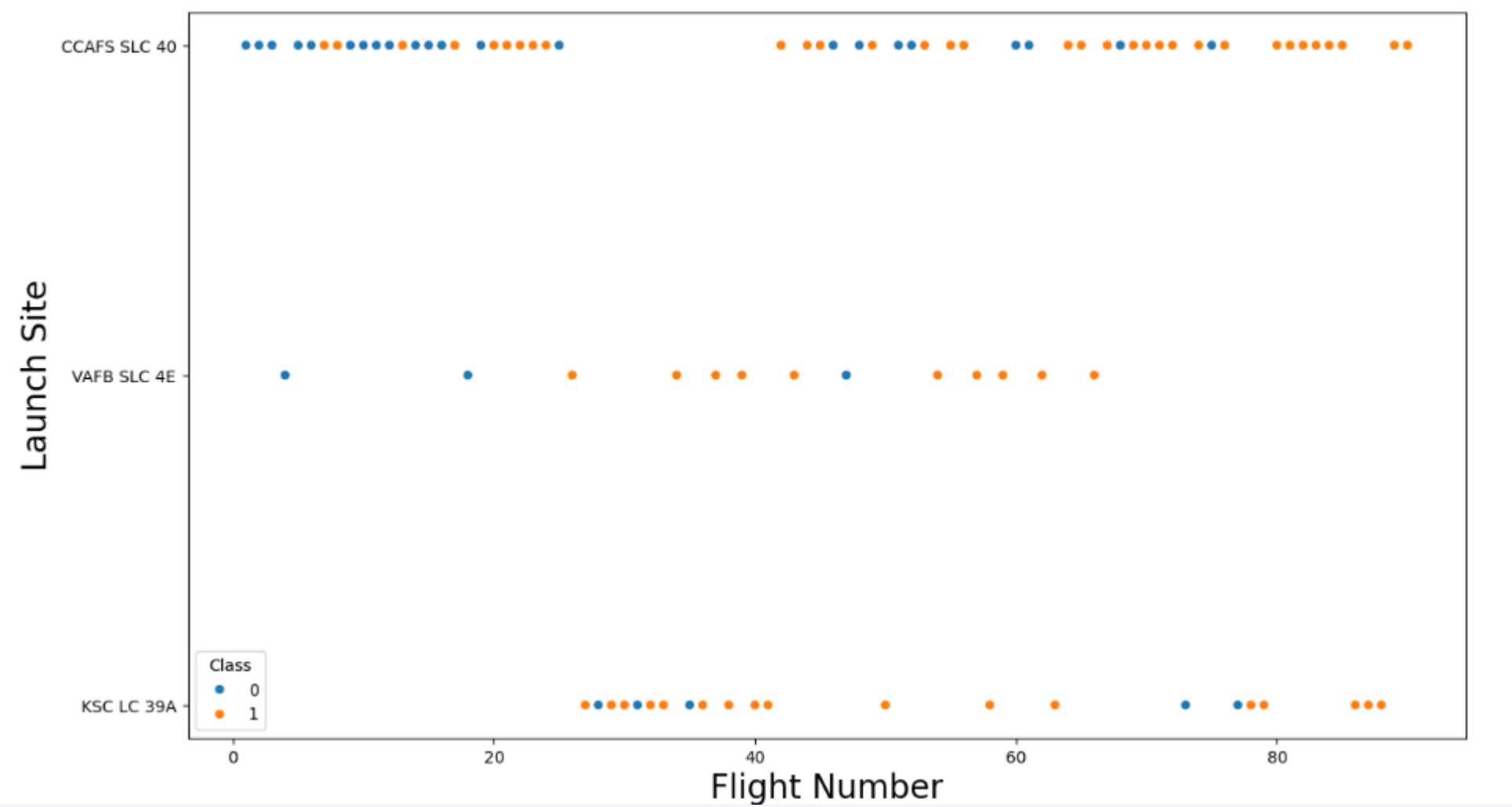
- The Scikit-learn library is used to create machine learning models for the space mission data. The machine learning prediction process includes:

1. Standardizing the data
2. Splitting the data into training and test sets
3. Creating machine learning models, including:
    1. Logistic regression
    2. Support vector machine (SVM)
    3. Decision tree
    4. K-nearest neighbors (KNN)
4. Training the models on the training data
5. Tuning the hyperparameters to find the best-performing models
6. Evaluating the models based on accuracy scores and confusion matrices

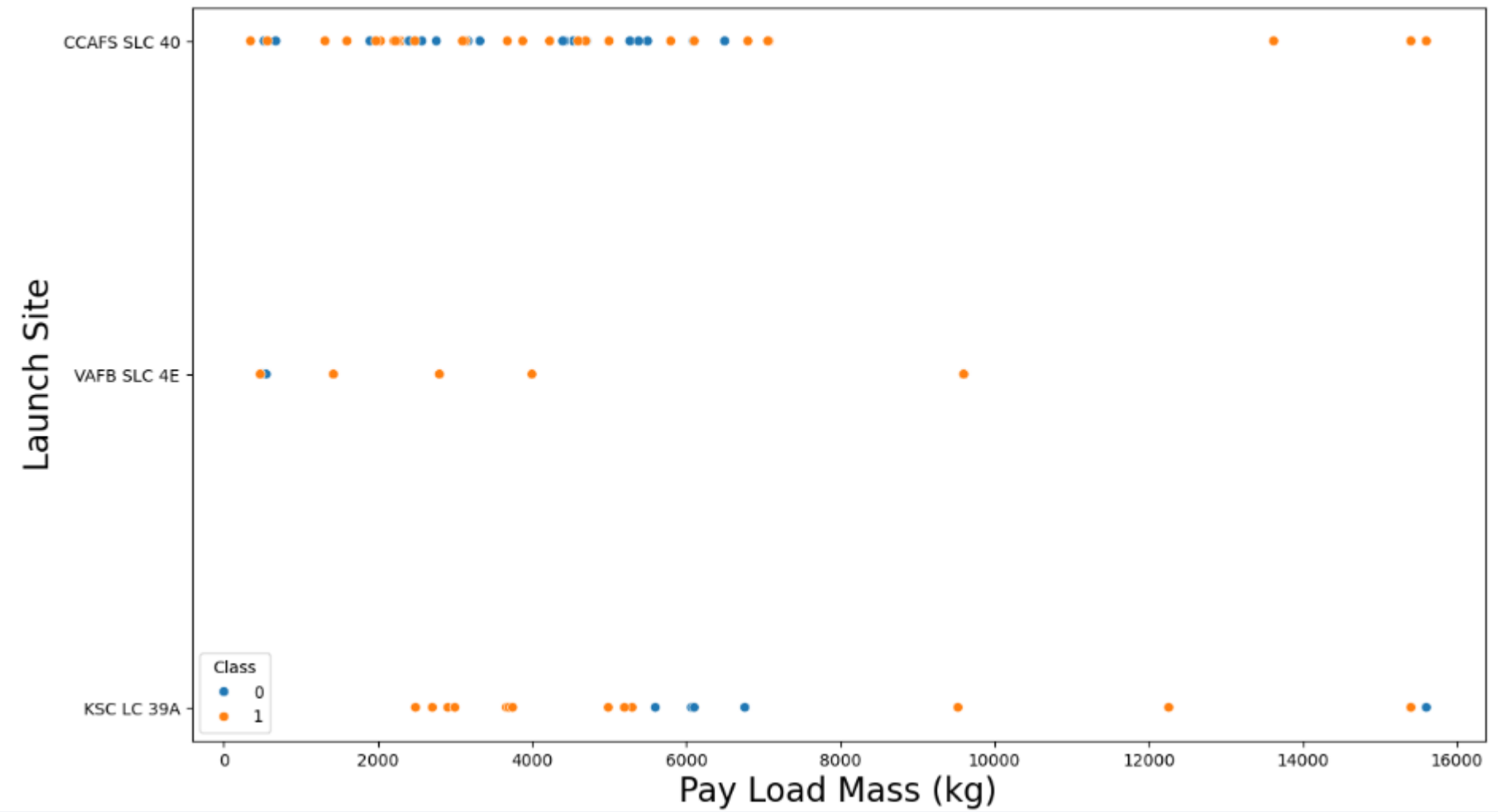
- The Scikit-learn library provides the necessary functions and tools to perform this end-to-end machine learning workflow on the space mission data.


Section 2
Insights drawn from EDA

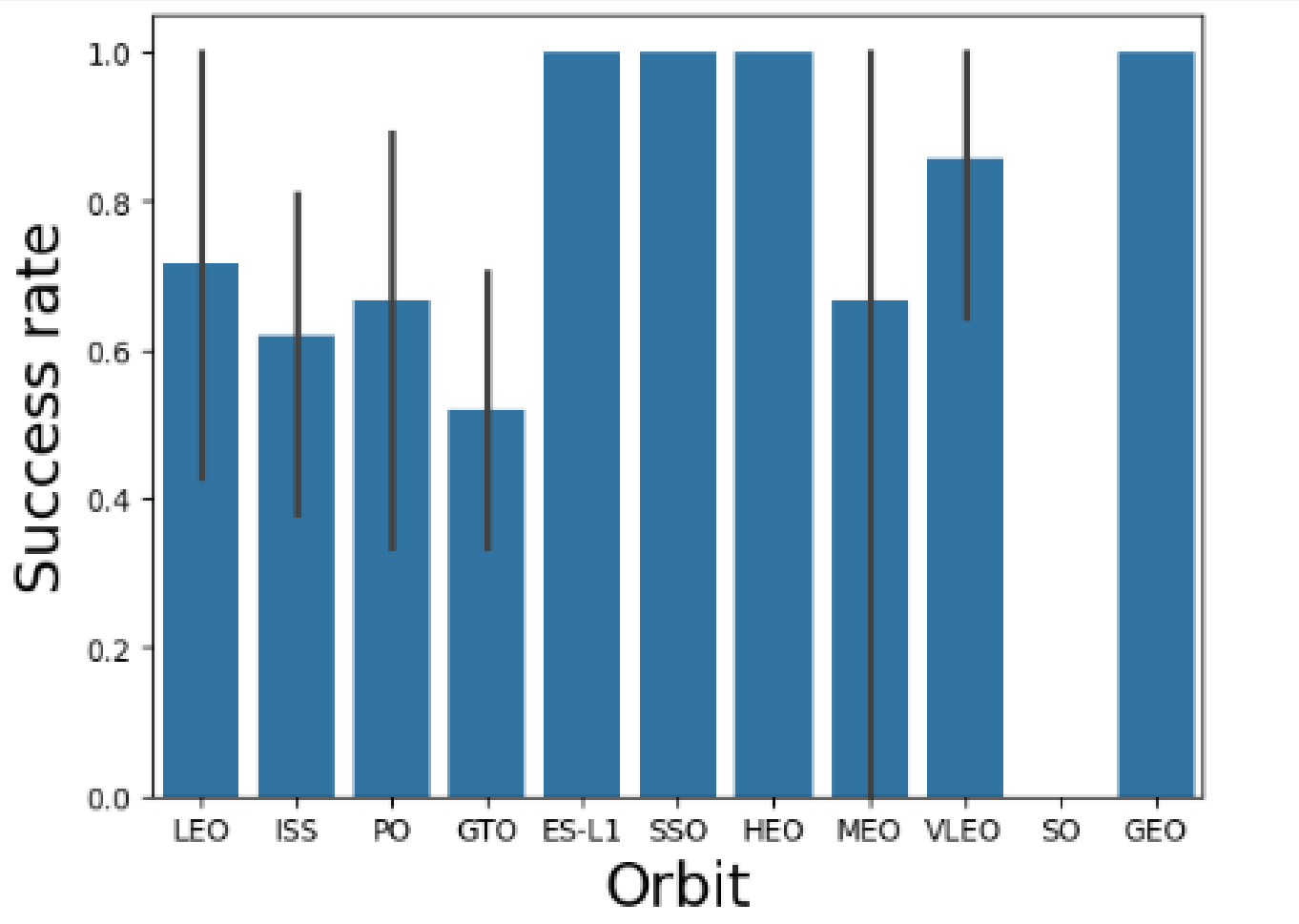
Flight Number vs. Launch Site



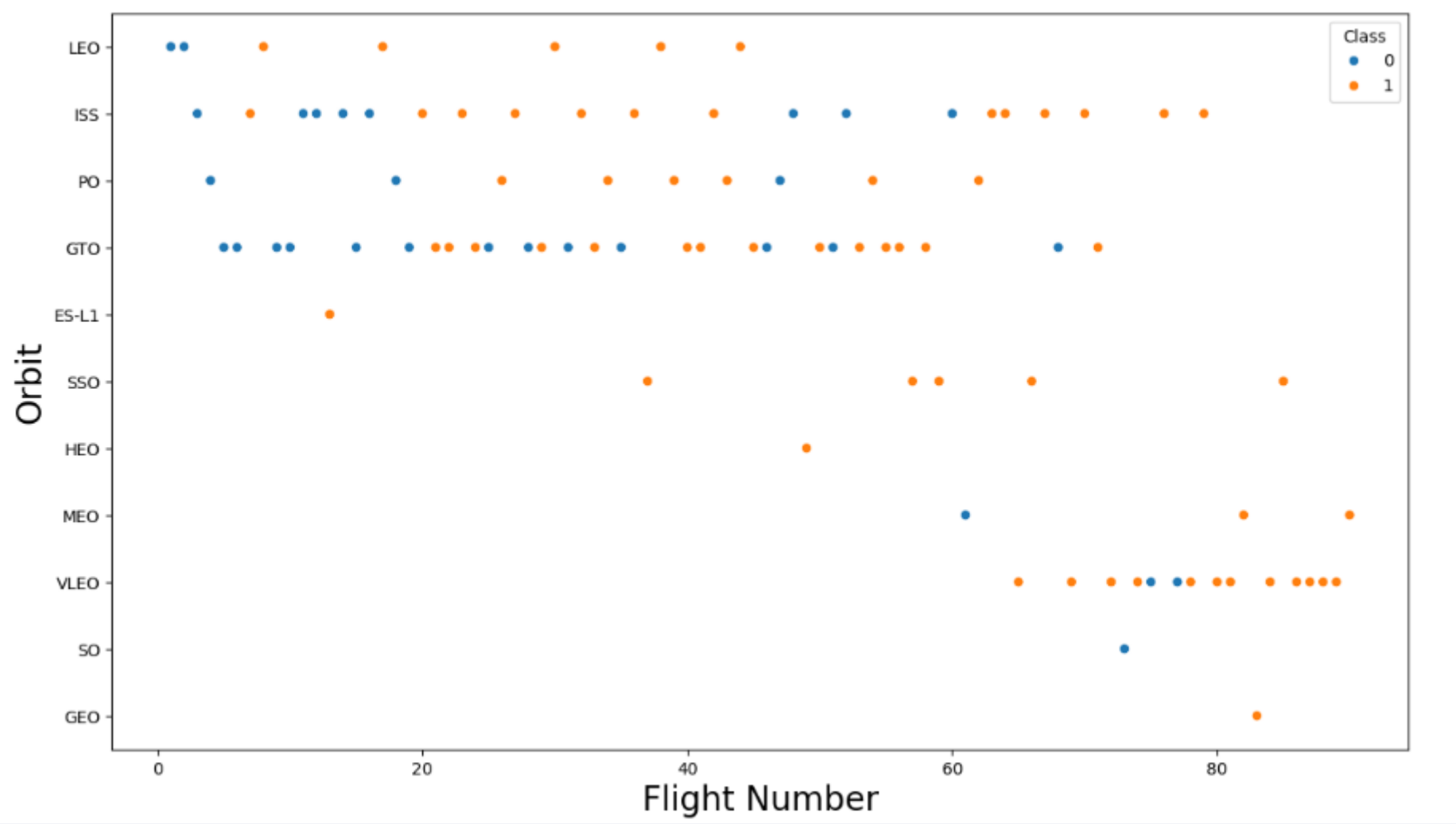
Payload vs. Launch Site



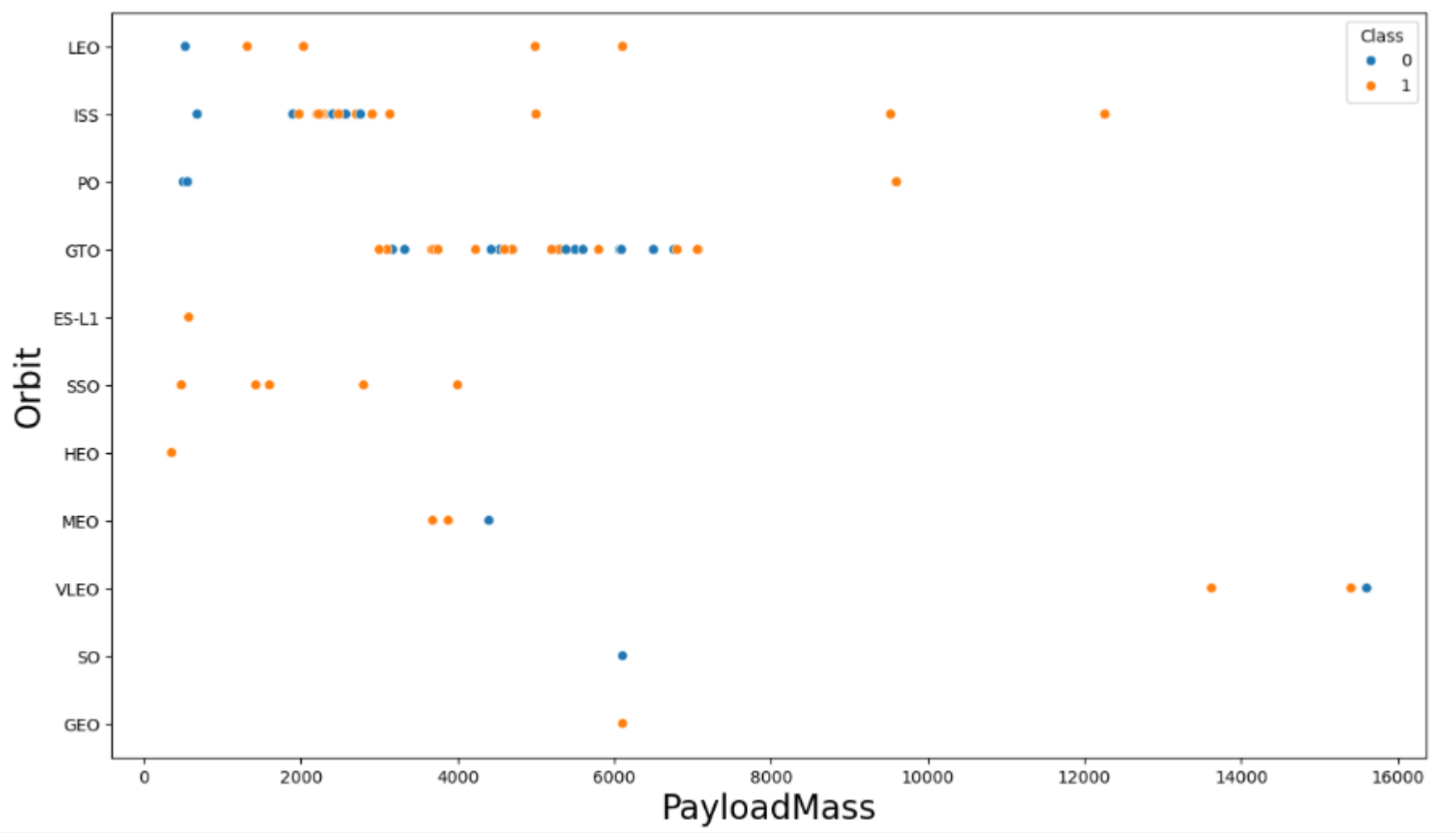
Success Rate vs. Orbit Type



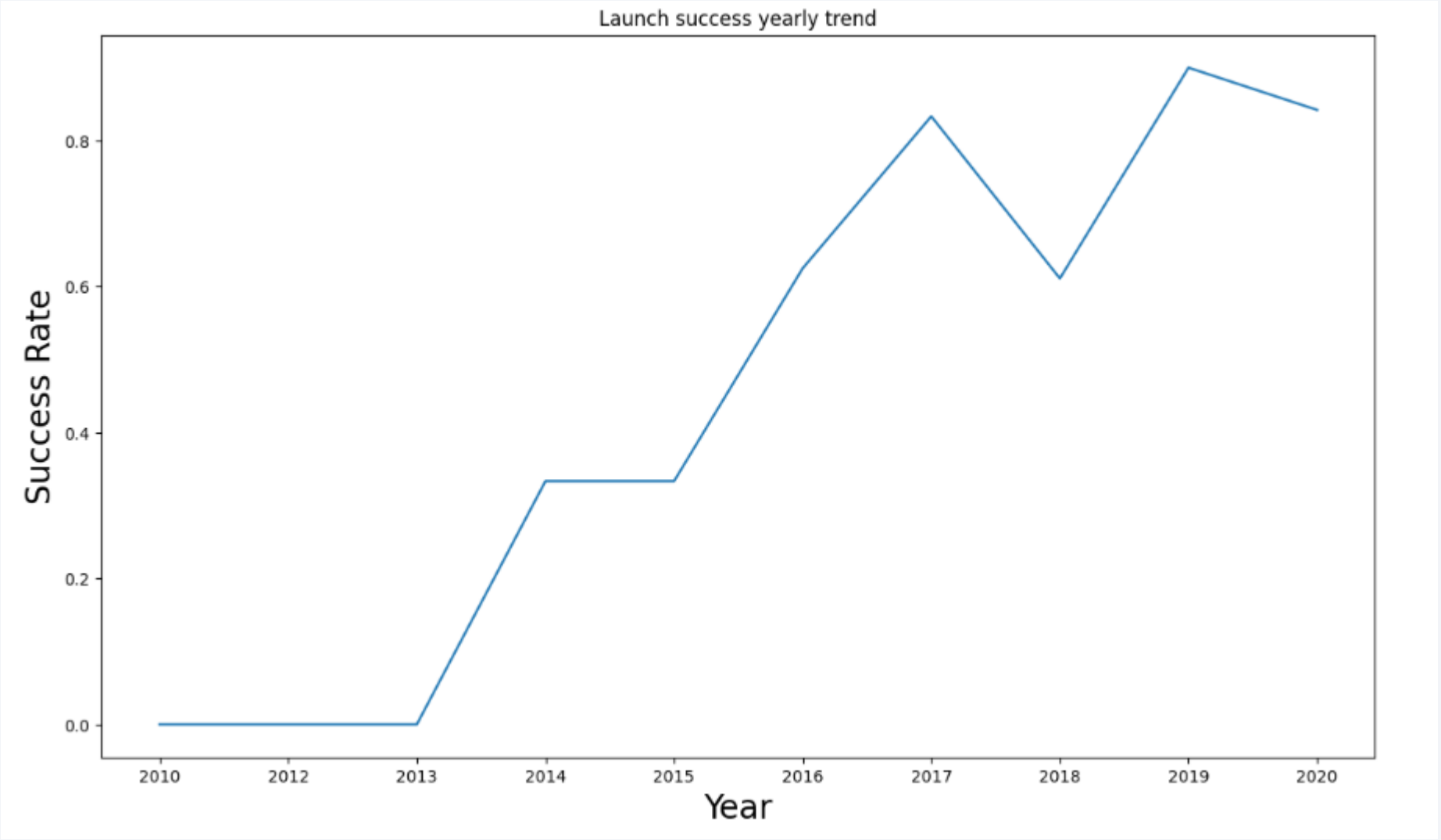
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

Launch_Sites
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Total payload mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

Average payload mass by Booster Version F9 v1.1
2928.4

First Successful Ground Landing Date

Date of first successful landing outcome in ground pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1

Boosters Carried Maximum Payload

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

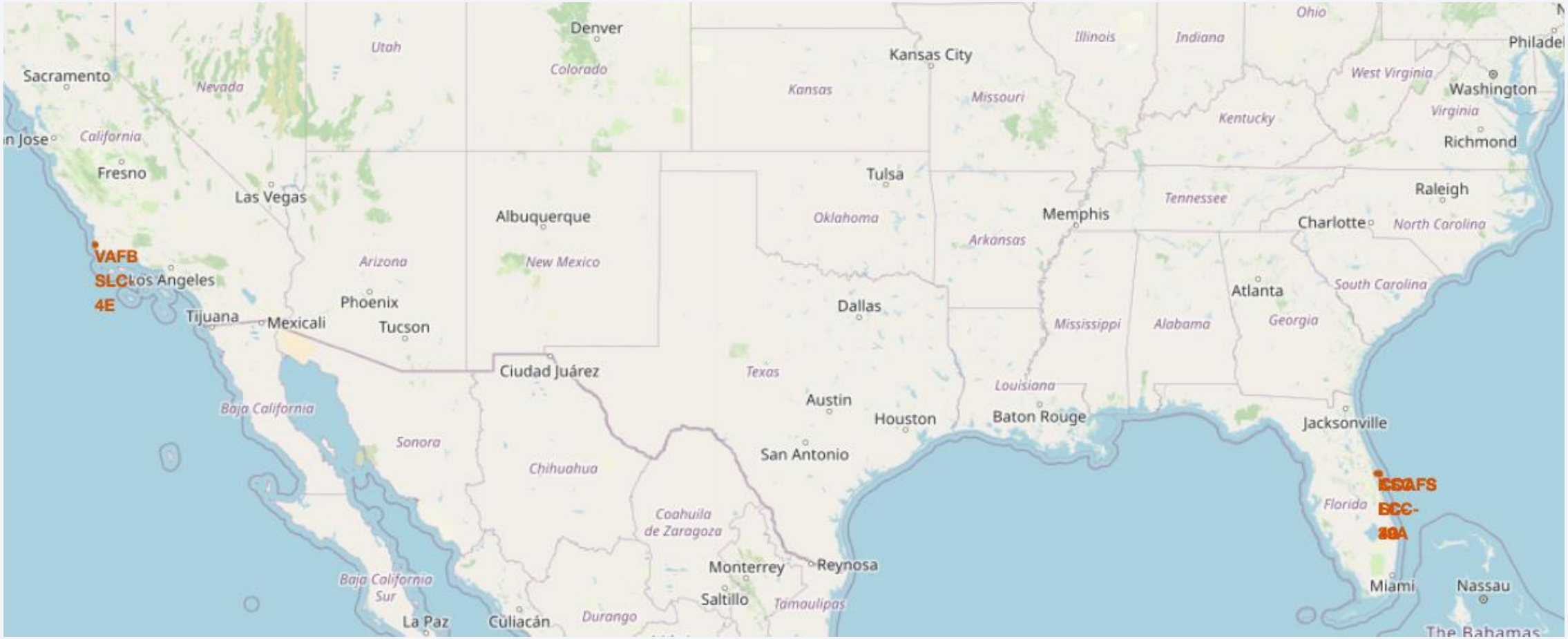
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

landing__outcome	landing_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

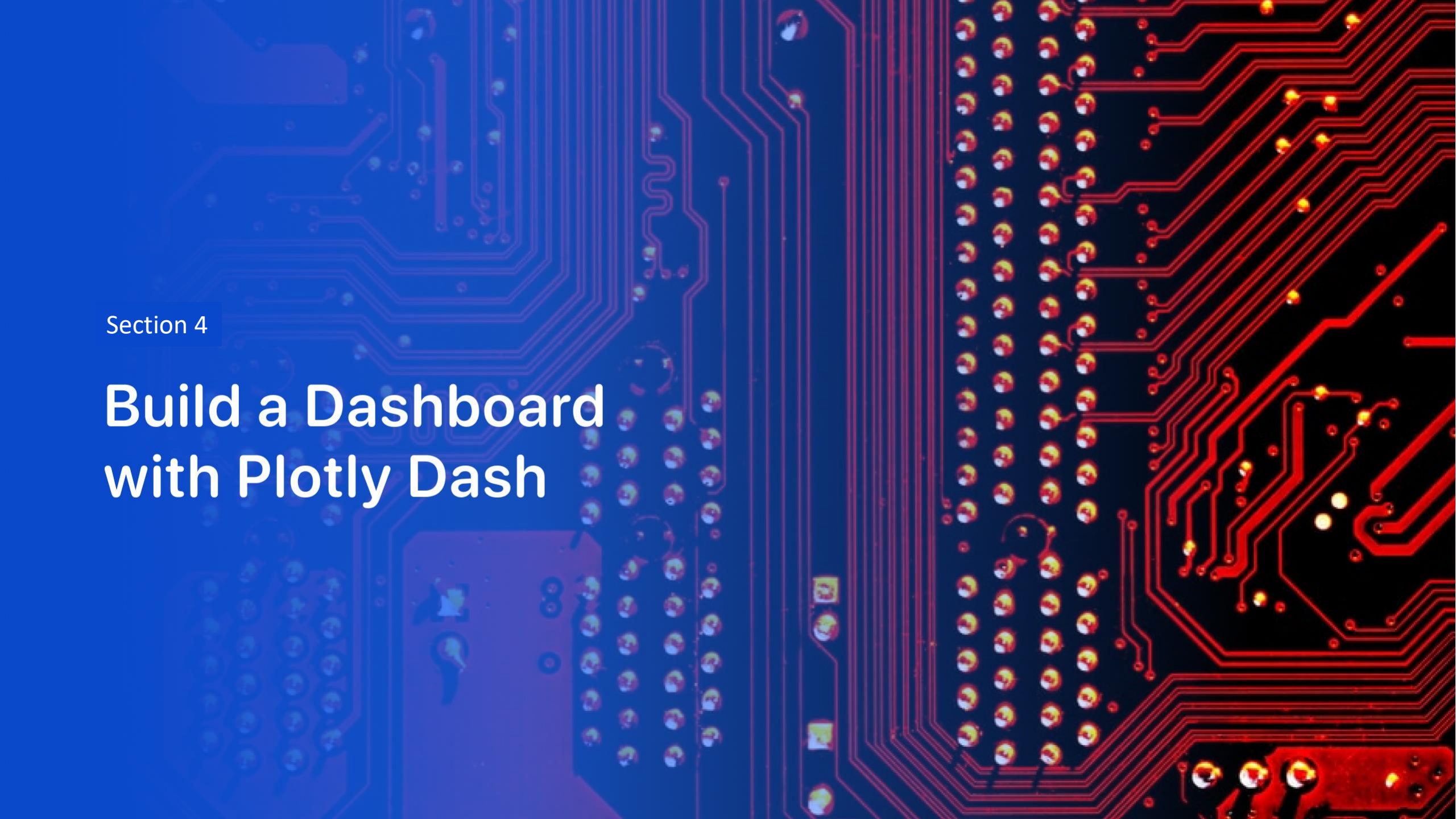
Launch Sites Location



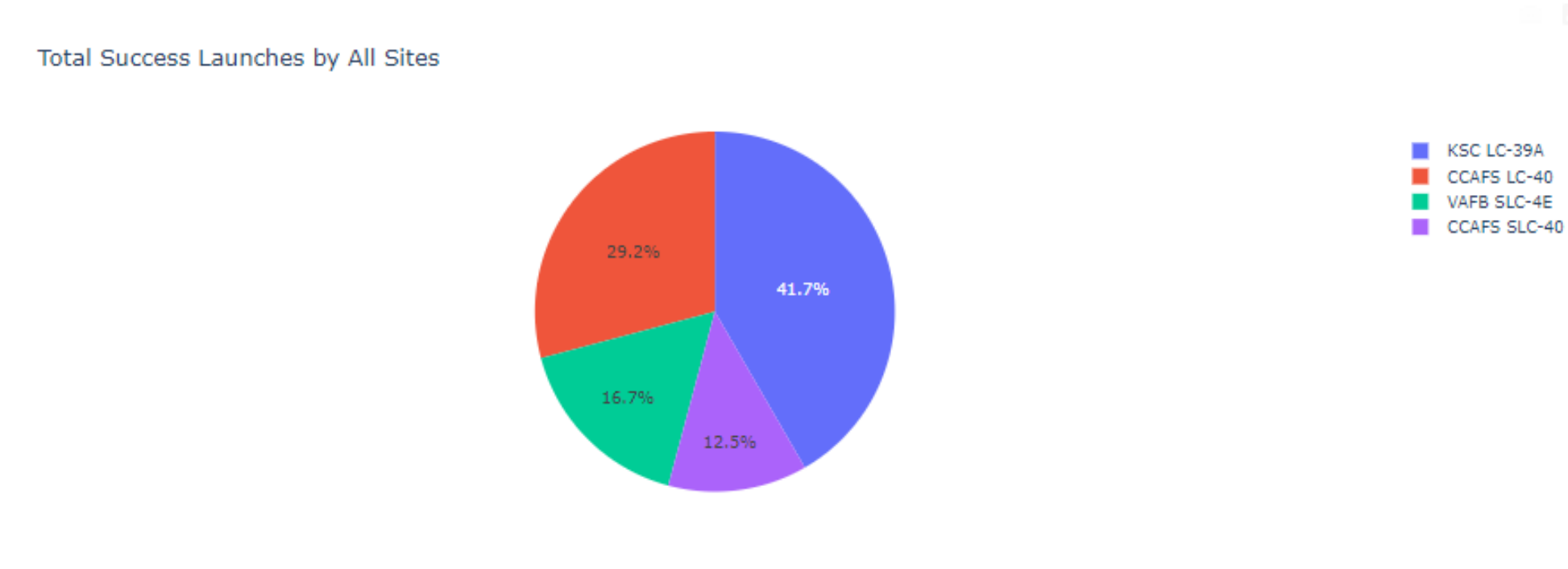
Launch Outcomes on Map

- Green tags represent successful launches while red tags represent failed Launches.

Distance Between a Launch Site to Proximities

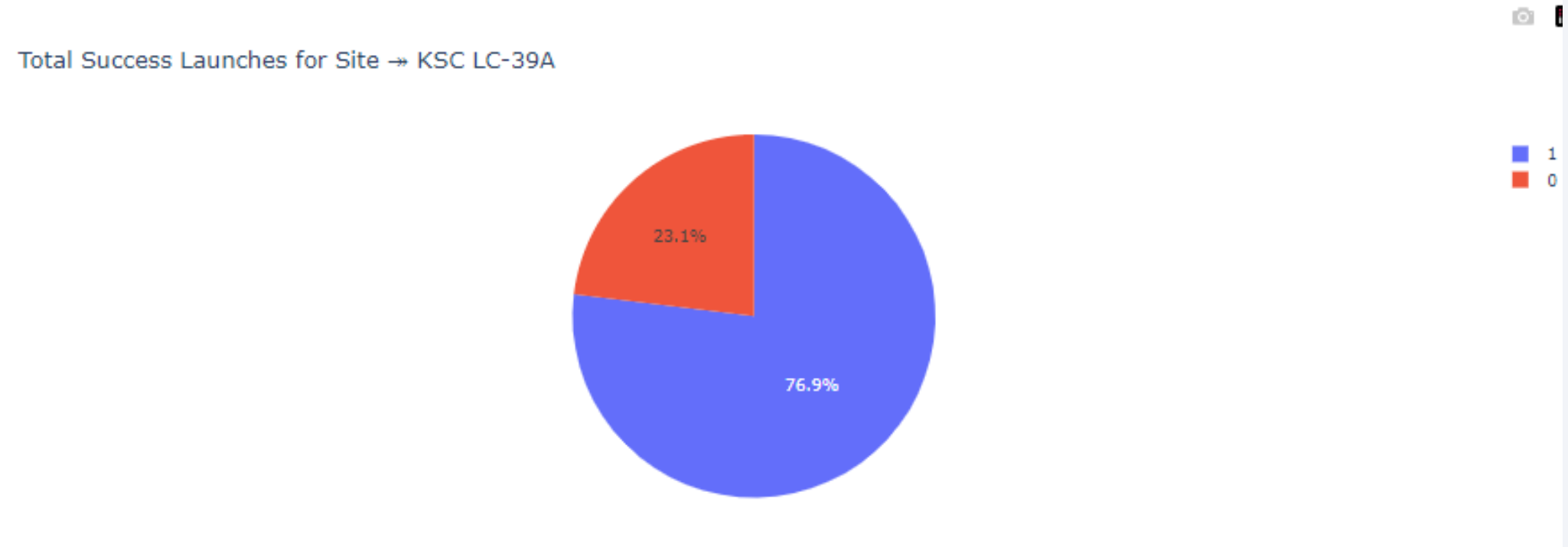

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by All Sites

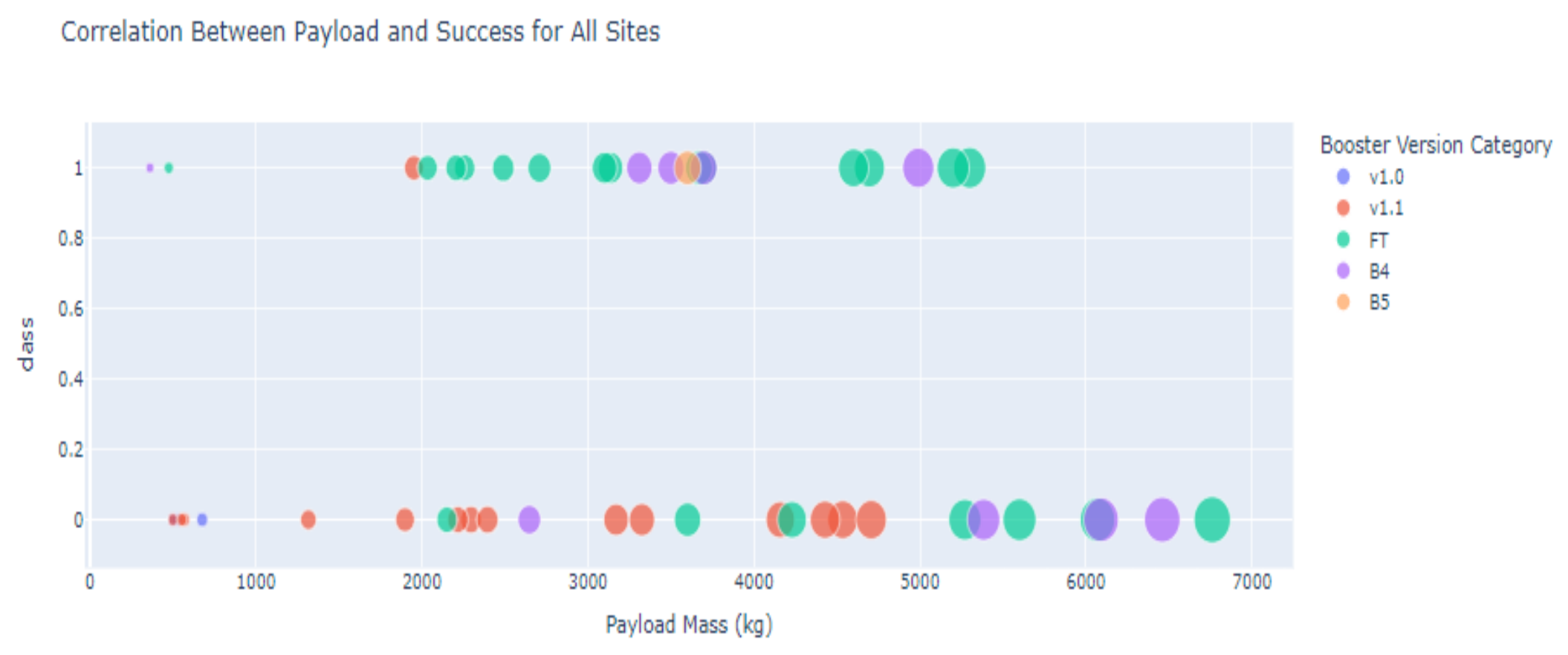


Launch Site with Highest Launch Success Ratio

- KSC LC-39A is the launch site with the highest launch success ratio.
- 0 represents failed launches while 1 represents successful launches.



Correlation Between Payload and Success for All sites

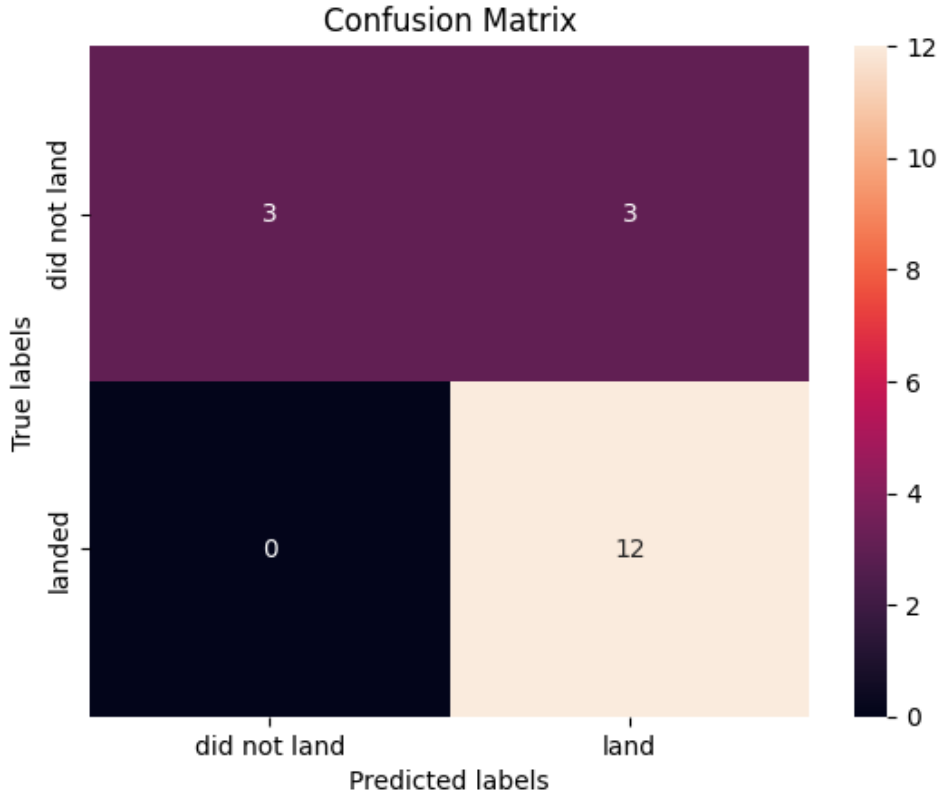


Section 5

Predictive Analysis (Classification)

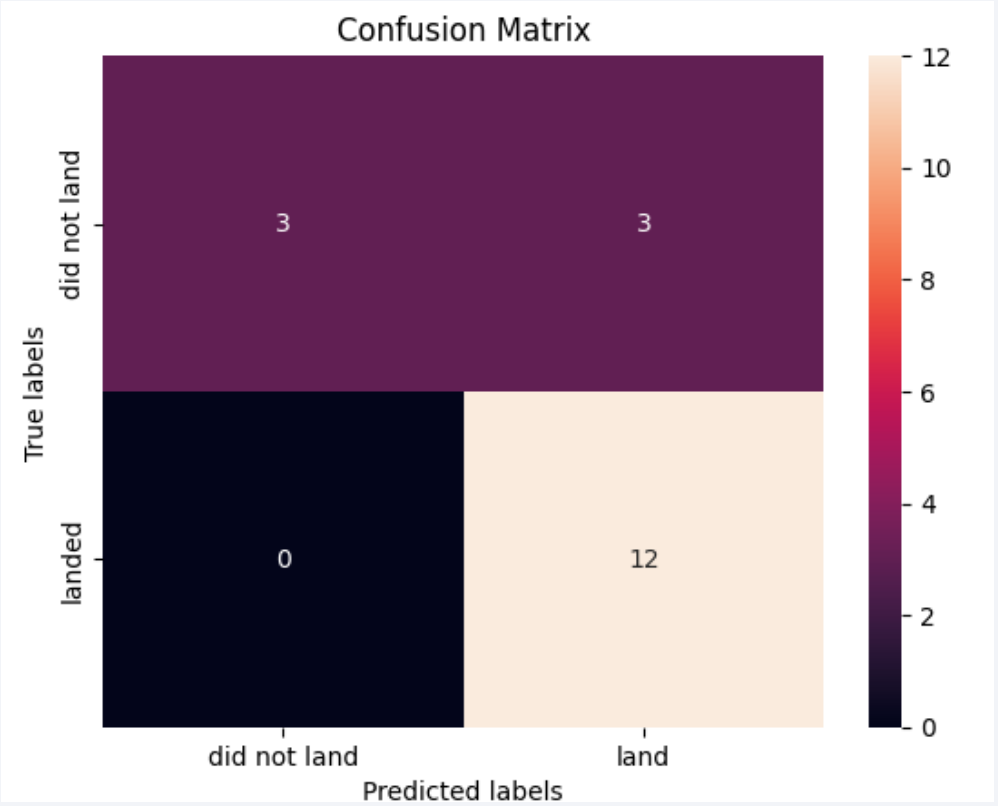
Logistic Regression

- GridSearchCV best scores: 0.8464
- Accuracy score: 0.8333
- Confusion Matric:



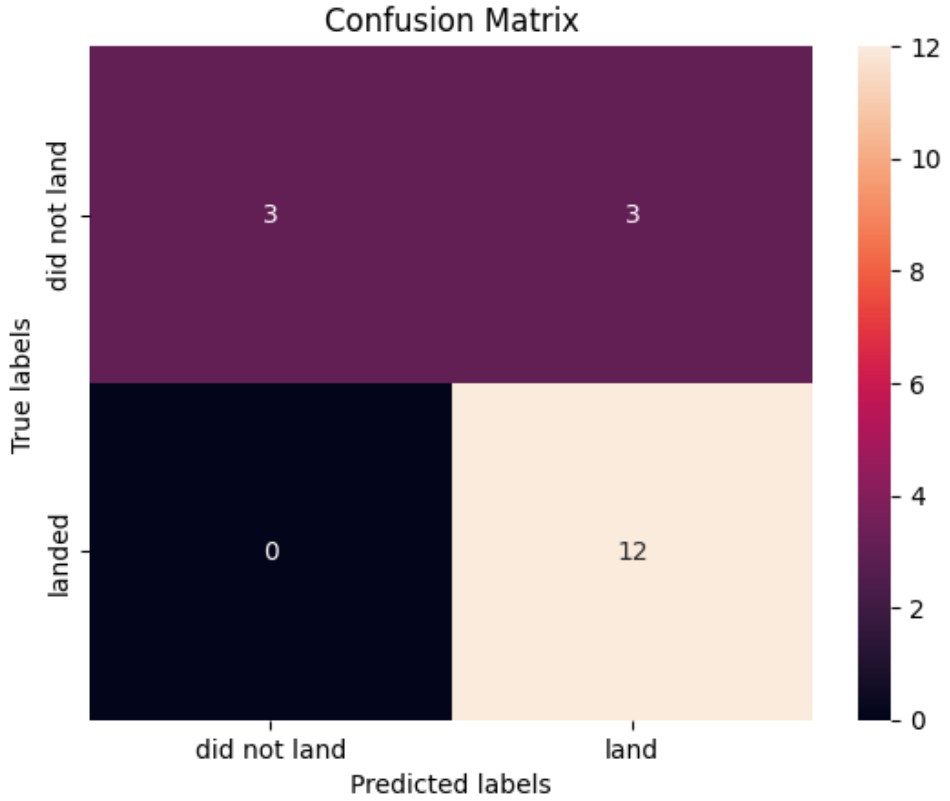
Support Vector Machine (SVM)

- GridSearchCV best scores: 0.8482
- Accuracy score: 0.8333
- Confusion Matric:



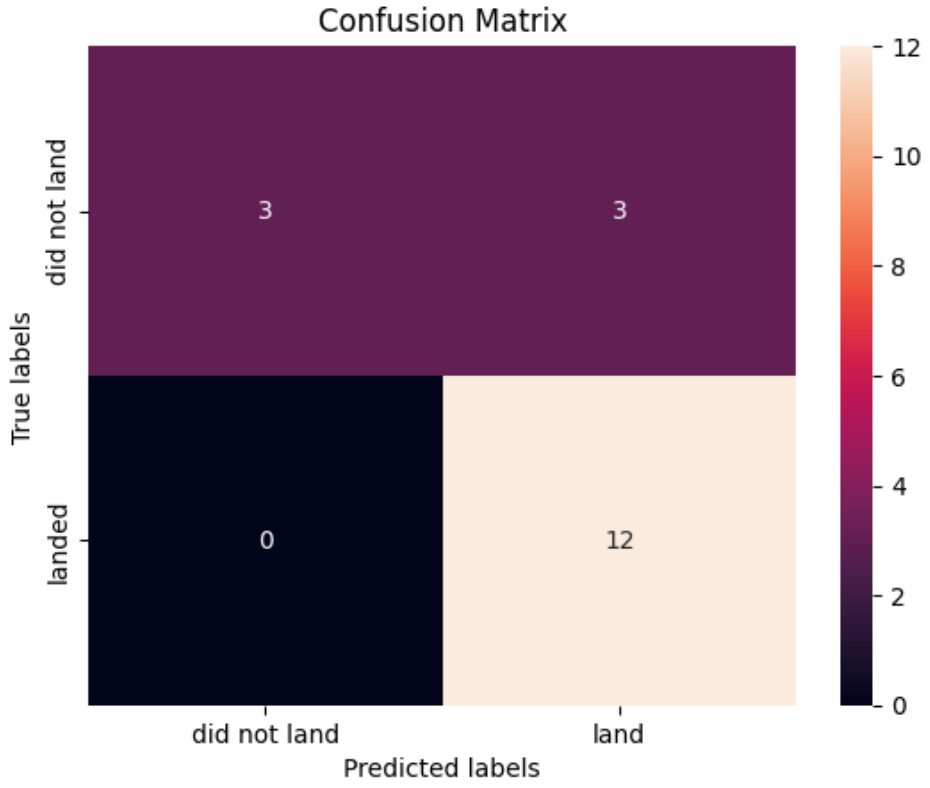
Decision Tree

- GridSearchCV best scores: 0.8643
- Accuracy score: 0.6667
- Confusion Matric:



K-nearest Neighbors (KNN)

- GridSearchCV best scores: 0.8482
- Accuracy score: 0.8333
- Confusion Matric:



Result

- Based on the GridSearchCV best score, decision tree would be the most appropriate model to be used in this case.

	Best scores
Logistic regresssion	0.846429
SVM	0.848214
Decision tree	0.864286
KNN	0.848214

Conclusions

- In this project, we aim to predict whether the first stage of a Falcon 9 launch will successfully land, in order to determine the cost of the launch.
- Each feature of a Falcon 9 launch, such as its payload mass or orbit type, may have a specific impact on the mission outcome.
- Multiple machine learning algorithms are used to learn the patterns from past Falcon 9 launch data, in order to create predictive models that can forecast the outcome of future Falcon 9 launches.
- Of the four machine learning algorithms employed, the predictive model generated by the decision tree algorithm demonstrated the best performance.

Thank you!